John von Neumann Institute for Computing

HANSWURST: Fast Efficient Multiple Protein Structure Alignments

Th. Margraf, A. Torda

published in

From Computational Biophysics to Systems Biology (CBSB08),
Proceedings of the NIC Workshop 2008,
Ulrich H. E. Hansmann, Jan H. Meinke, Sandipan Mohanty,
Walter Nadler, Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **40**, ISBN 978-3-9810843-6-8, pp. 313-316, 2008.

http://www.fz-juelich.de/nic-series/volume40

HANSWURST: Fast Efficient Multiple Protein Structure Alignments

Thomas Margraf and Andrew Torda

Center for Bioinformatics, University of Hamburg, Bundesstrasse 43, 20146 Hamburg, Germany
E-mail: {*margraf, torda*}*@zbh.uni-hamburg.de*

We have built a multiple structure alignment tool which is able to compute alignments and phylogenies of vast numbers of proteins. HANSWURST is a progressive alignment method with time complexity in the class of O(n^2). It takes advantage of a probabilistic representation of protein structure which allows for the calculation of meaningful average representations of clusters of proteins, and the alignment of those representations. Our tool scales to well over 1000 structures which is enough to cover even the largest protein families.

1 Introduction

One or two homologous sequences whisper [...]; a full multiple alignment shouts out loud.[1] This quote very eloquently describes the usefulness of multiple alignments. The significance of matches in pairwise alignments can be difficult to judge against the background noise of random matches. In multiple alignments however, random matches across a reasonable number of structures are so improbable that there is little room for doubts about their significance. This is doubly true for multiple structure alignments which really begin to shine when the relationships between proteins become so remote that sequence methods start to break down.

Common application areas for multiple structure alignments are in homology modeling[2], protein function prediction[3], creation of substitution matrices[4], phylogeny[5] and structure classification[6].

HANSWURST is built on the assumption that local interactions between atoms are the most important factor in determining the overall structure of a protein. Therefore, long stretches of high local similarity should also lead to high global similarity. From this reasoning follows that HANSWURST's aim is not to produce alignments with optimal global similarity scores such as RMSD. Instead, good global scores are considered to be a property which emerges from local similarity.

This is almost the exact opposite of the ideas behind traditional multiple structure alignment methods which sacrifice sensitivity for lower structural alignment scores.

2 Materials and Methods

This work builds on many existing methods such as AutoClass[7], rigid body superposition[8], dynamic-programming sequence alignment algorithms[9], hierarchical clustering algorithms[10], multiple sequence alignment methods and the computation of consensus probability vectors by averaging.

The basis of the alignment method is a bayesian classification of protein structure fragments using the AutoClass program[7,11].

Based on the class descriptions in such a classification, we can calculate the probability of a given protein fragment being in a certain class. The set of all class membership probabilities for a given protein fragment can be represented as a probability vector.

The dot product of two such vectors can be used as a similarity measure between two peptide fragments. This score can then be used instead of a substitution matrix in standard sequence alignment methods[9].

The resulting pairwise alignments of all vs. all structures one wishes to align are then used to fill a distance matrix. On the basis of this matrix, various clustering algorithms can be used to construct a guide tree. Currently, the best such algorithm is derived from the UPGMA method[10] and uses alignments of average probability vectors to estimate the distances between internal nodes of the guide tree. Such consensus probability vectors are computed by averaging the class membership probabilities of each fragment in a given column in the alignment. Gaps have no class memberships and thus do not contribute to the average. This concept allows each node in the guide tree to have a set of probability vectors associated with it which represent the average class memberships of that nodes descendants. Since all the information required to compute a pairwise alignment is avail-

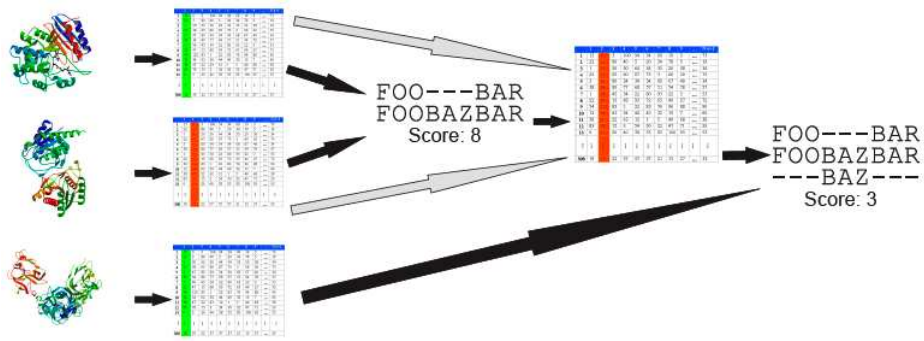

Figure 1. Illustration of the progressive construction of a three way alignment.

able for any cluster of structures, distances between clusters can be calculated by aligning their associated probability vectors. This removes the need to estimate distances during the construction of the guide tree and therefore improves its quality. The alignments of those average probability vectors are also used to merge the pairwise alignments according to the guide tree.

3 Results and Discussion

As a demonstration of our method's capabilities, we took 818 proteins with pairwise sequence identities below 25% and built a multiple structure alignment. Computing this alignment took just over 8 hours of CPU time. Anecdotal evidence suggests that this method matches almost three times as many residues as competing methods[8, 9] with some increase in the RMSD scores of the resulting superpositions. The improvement of the consensus clustering method over traditional clustering methods can be regarded as the

biggest advantage of HANSWURST over competing structural alignment methods. By representing protein structures as sets of probability vectors with regard to a fixed classification, one can easily calculate characteristic representations of clusters of proteins by averaging class membership probabilities of aligned residues in a cluster.

In combination with the method's speed and scalability, this enables the creation of multiple structure alignments of vast numbers of distantly related proteins.

References

1. T J Hubbard, A M Lesk, and A Tramontano, *Gathering them in to the fold.*, Nat Struct Biol, **3**, no. 4, 313, 1996.
2. Anthony J Russell and Andrew E Torda, *Protein sequence threading: Averaging over structures.*, Proteins, **47**, no. 4, 496–505, 2002.
3. Roman A Laskowski, James D Watson, and Janet M Thornton, *ProFunc: a server for predicting protein function from 3D structure.*, Nucleic Acids Res, **33**, no. Web Server issue, W89–93, 2005.
4. Manoj Tyagi, Venkataraman S Gowri, Narayanaswamy Srinivasan, Alexandre G de Brevern, and Bernard Offmann, *A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications.*, Proteins, **65**, no. 1, 32–39, 2006.
5. D F Feng and R F Doolittle, *Progressive sequence alignment as a prerequisite to correct phylogenetic trees.*, J Mol Evol, **25**, no. 4, 351–360, 1987.
6. L Holm and C Sander, *Dali/FSSP classification of three-dimensional protein folds.*, Nucleic Acids Res, **25**, no. 1, 231–234, 1997.
7. P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, *AutoClass: A Bayesian Classification System*, in: Proceedings of the Fifth International Conference on Machine Learning, pp. 54–64, Morgan Kaufmann. 1988.
8. R. Diamond, *A note on the Rotational Superposition Problem*, Acta Cryst., **A**, no. 44, 211–216, 1988.
9. SB Needleman and CD Wunsch, *A general method applicable to the search for similarities in the amino acid sequence of two proteins*, Journal of Molecular Biology, **48**, 443–453, 1970.
10. R.R. Sokal and C.D. Michener, *A statistical method for evaluating systematic relationships*, University of Kansas Scientific Bulletin, **28**, 1958.
11. G Schenk, T Margraf, and AE Torda, *Protein sequence and structure alignments within one framework.*, Algorithms Mol Biol, **3**, no. 1, 4, 2008.